Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila

Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of *Drosophila*

Daniel L. Halligan,[1] Adam Eyre-Walker,[2] Peter Andolfatto,[1,3] and Peter D. Keightley[1,4]

[1]*University of Edinburgh, School of Biological Sciences, Edinburgh EH9 3JT, UK;* [2]*University of Sussex, Centre for the Study of Evolution and School of Biological Sciences, Brighton BN1 9QG, UK*

We develop methods to infer levels of evolutionary constraints in the genome by comparing rates of nucleotide substitution in noncoding DNA with rates predicted from rates of synonymous site evolution in adjacent genes or other putatively neutrally evolving sites, while accounting for differences in base composition. We apply the methods to estimate levels of constraint in noncoding DNA of *Drosophila*. In introns, constraint (the estimated fraction of mutations that are selectively eliminated) is absolute at the 5′ and 3′ splice junction dinucleotides, and averages 72% in base pairs 3–6 at the 5′-end. Constraint at the 5′ base pairs 3–6 is significantly lower in the lineage leading to *Drosophila melanogaster* than in *Drosophila simulans*, a finding that agrees with other features of genome evolution in *Drosophila* and indicates that the effect of selection on intron function has been weaker in the *melanogaster* lineage. Elsewhere in intron sequences, the rate of nucleotide substitution is significantly higher than at synonymous sites. By using intronic sites outside splice control regions as a putative neutrally evolving standard, constraint in the 500 bp of intergenic DNA upstream and downstream regions of protein-coding genes averages ~44%. Although the estimated level of constraint in intergenic regions close to genes is only about one-half of that of amino acid sites, selection against single-nucleotide mutations in intergenic DNA makes a substantial contribution to the mutation load in *Drosophila*.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AY459538–AY459582.]

Understanding the functional significance of intronic and intergenic noncoding DNA sequences is one of the major challenges in genomics research at present. If functional elements of the genome are close to adaptive optima owing to past directional selection, these sequences are expected to show evidence of purifying selection. This manifests itself as a lower rate of between-species nucleotide substitution when comparisons are made with evolutionary rates in neutrally evolving DNA segments having similar base composition and mutation rates. The level of functional conservation in the genome is important in determining the genome-wide mutation load due to the selective elimination of deleterious mutations (Kondrashov 1995), and this affects several important evolutionary issues (Charlesworth and Charlesworth 1998). Although it is well established that most protein-coding sequences are strongly constrained, that is, that most amino acid altering mutations are deleterious and become selectively eliminated (e.g., Li 1997), functional conservation in noncoding DNA has been much less well studied and is subject to controversy. Although some introns contain regulatory elements, several comparative studies suggest that introns evolve largely free from selective constraints (Gilbert 1978; Li and Graur 1991; Li 1997). However, recent genome-wide interspecific comparisons imply that intron sequences are subject to significant evolutionary pressures (Jareborg et al. 1999; Shabalina and Kondrashov 1999; Bergman and Kreitman 2001). In comparisons involving mammals, the issue of relative rates of substitution is complicated by the presence of methylated CG dinucleotides, which have greatly elevated mutation rates, and whose frequency varies between coding and noncoding DNA and between different categories of noncoding DNA (Chen and Li 2001; Hellmann et al. 2003; Subramanian and Kumar 2003). In intergenic DNA, genome-wide interspecific comparisons (Jareborg et al. 1999; Shabalina and Kondrashov 1999; Bergman and Kreitman 2001; Shabalina et al. 2001), and comparisons of known or putative regulatory elements (Ludwig and Kreitman 1995; Glazko et al. 2003) have also revealed substantial constraints, but the overall level of conservation and the distribution of conserved elements in intergenic regions of the genome is still largely unknown.

Present methods to quantify functional constraints in DNA sequences mostly depend on comparative genomics approaches. They relate to a method for inferring the genome-wide deleterious mutation rate based on sequence divergence (Kondrashov and Crow 1993). Shabalina and Kondrashov (1999) proposed that the proportion of bases that are subject to strong purifying selection can be quantified by comparing the genomes of distantly related species. It is assumed that homologous segments lacking similarity are saturated with nucleotide and/or indel substitutions, and are evolving free from functional constraint, whereas segments showing similarity ("hits") are under strong functional constraint. Constraint is quantified as the fraction of conserved nucleotides in the hits, which is assumed to comprise bases under strong purifying selection. Potential difficulties with the approach are variation across the genome in the mutation rate, which could make nonfunctional elements appear functional (Clark 2001), and obtaining the correct (or most probable) sequence alignment. If the DNA sequence alignment method is heuristic, and, for example, genuine similarities are missed, then functional elements could appear nonfunctional.

A second general approach for quantifying evolutionary constraint also uses comparisons between DNA segments of re-

[3]Present address: Department of Zoology, University of Toronto, Toronto, ON M5S 3G5, Canada.
[4]Corresponding author.
E-MAIL Peter.Keightley@ed.ac.uk; FAX 44 131 650 5443.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1329204.

lated species, but uses sequences from species showing lower levels of divergence that are far from saturation. It is based on comparing the rate of evolution of a putative functional segment of noncoding DNA with the rate of evolution of a DNA segment or a category of nucleotide sites that is assumed to be evolving free from constraint (a neutral segment), that has the same mutation rate, and can therefore act as a standard. Constraint is the factor by which evolution is slowed down in the functional segment (Kimura 1983). Nucleotides are assumed to fall into two classes in the functional sequence: neutral, which evolve at the same rate as the neutral sequence; or strongly constrained, in which mutations are eliminated unconditionally by natural selection. The neutral segment should be adjacent to the functional segment, thereby making the assumption of equality of mutation rates defensible. A close to ideal situation would be to compare the rate of evolution of a pseudogene (assumed to completely lack function) to that of an adjoining noncoding DNA segment. Unfortunately, because in many taxa, including *Drosophila*, pseudogenes are uncommon (Petrov et al. 1996), an alternative category of neutrally evolving sequences is needed. A candidate for such a category is synonymous sites of genes, because changes in these do not lead to change in the amino acid sequence. In many taxa, including *Drosophila*, however, there is evidence of past selection acting on synonymous codon usage (Shields et al. 1988), and this could retard rates of evolution at synonymous sites (Li 1997, Chapter 7).

In this paper, our initial approach is to use synonymous sites of *Drosophila* protein-coding genes as a standard for estimation of constraint in adjacent noncoding DNA. Features of the dynamics of synonymous site evolution indicate that selection at synonymous sites is weak or, in some cases, absent in *Drosophila*. In the lineage leading to *Drosophila melanogaster* from its common ancestor with *Drosophila simulans*, there has been a surge in the rate of preferred to unpreferred synonymous substitutions (Akashi 1995, 1996; McVean and Vieira 2001). This surge in the rate of substitution indicates a genome-wide relaxation of selection at synonymous sites, possibly because of demographic changes that have changed the efficiency of natural selection. In *D. melanogaster*, a population genetics analysis of the pattern of synonymous site divergence indicates that selection has been relaxed to the point of being completely absent (McVean and Vieira 2001). Further evidence for low levels of selection presently acting on *D. melanogaster* synonymous codon usage comes from an analysis of the frequency spectrum of segregating synonymous sites (Akashi 1999). A weakening of selection to approximately one-half of that in the ancestral species is estimated to have occurred in the *D. simulans* lineage (McVean and Vieira 2001). Furthermore, weak selection of the magnitude thought to be acting on synonymous codon usage in *Drosophila* (Akashi 1995, 1996) is predicted to have only a small effect on substitution rates (Eyre-Walker and Bulmer 1995). Recently, an apparent excess of preferred to unpreferred synonymous site substitutions has been reported in the *Xdh* gene of several *Drosophila* species (Begun and Whitley 2002). Possible explanations for this observation are an evolutionary shift in base composition towards A/T nucleotides in many *Drosophila* lineages (the explanation favored by Begun and Whitley 2002; see also Duret et al. 2002), a general weakening of selection in *Drosophila* lineages, or an artifact of parsimony if nucleotide mutation rates are sufficiently variable.

There are two additional difficulties in interpreting comparisons between synonymous site or other putatively neutral site divergence and nucleotide divergence in noncoding DNA. First, differences in the rate of substitution can be induced by differences in base composition; this stems from variation in average mutation rates between different kinds of nucleotides. We address this by comparing expected and observed numbers of substitutions; expected numbers in a noncoding segment are predicted on the basis of substitution rates at synonymous or other putatively neutral sites of adjacent genes, after the compositional effect has been accounted for.

A second potential problem in analyzing evolutionary rates in noncoding DNA concerns inference of the correct sequence alignment. Consider two alternative plausible alignments of a pair of sequences containing at least one gap:

Alignment 1	Alignment 2
Three substitutions	One substitution
ATGCATGCG	ATGCATGCG
AT--CAGCA	AT-CA-GCA

If alignment 1 were taken as the true alignment, the fraction of nucleotide differences ($k = 3/7$) would be radically different from taking alignment 2 ($k = 1/7$). The uncertainty is due to the unknown pattern of indels (gaps) between the sequences. In general, putative alignments containing too many gaps relative to the true alignment tend to have too few nucleotide substitutions or vice versa, and the bias can be serious. A solution to this problem has been proposed by Thorne et al. (1991, 1992), who developed an algorithm to compute probabilities of alternative alignments according to explicit models of indel evolution. Here, we use a Monte Carlo approach, MCALIGN, to tackle the problem of aligning noncoding DNA (P.D. Keightley and T. Johnson, unpubl.). Noncoding DNA sequences are aligned according to a model of indel evolution that is parameterized by relative rates of indels and nucleotide substitutions in noncoding DNA of closely related *Drosophila* species, and the most probable alignment is used in the subsequent analysis.

RESULTS

Indel Frequencies in Noncoding DNA Between *D. simulans* and *D. sechellia* and Parameterization of Alignment Models

To investigate the frequency distribution of indels, and to parameterize models of indel evolution suitable for aligning *Drosophila* noncoding DNA, we compiled intronic and intergenic DNA sequences from homologous loci of *D. simulans* and *D. sechellia* (Table 1). These species were chosen because they are part of the *melanogaster* subgroup, and are sufficiently closely related as to make alignment of noncoding DNA by standard heuristic methods virtually unambiguous. The frequency distribution of indel length in three DNA categories is shown in Figure 1 (two long intronic indels of length 29 and 37 are omitted to aid clarity). Distributions of indel length are not dissimilar to geometric distributions, as has been suggested previously (Gu and Li 1995). Numbers of substitutions do not differ significantly between DNA categories: A likelihood ratio test for heterogeneity among indel rates relative to substitution rates is nonsignificant (Table 1;

Table 1. Rates of Nucleotide Substitution (k) and Relative Rates of Indels (θ) in Noncoding DNA of *D. simulans* and *D. sechellia*

DNA category	No. of loci	Total no. bp	Substitutions	k	Indels	θ
Intronic	24	6302	193	0.0306	44	0.228
5′ intergenic	15	3094	85	0.0275	9	0.106
3′ intergenic	14	3159	101	0.0320	18	0.178

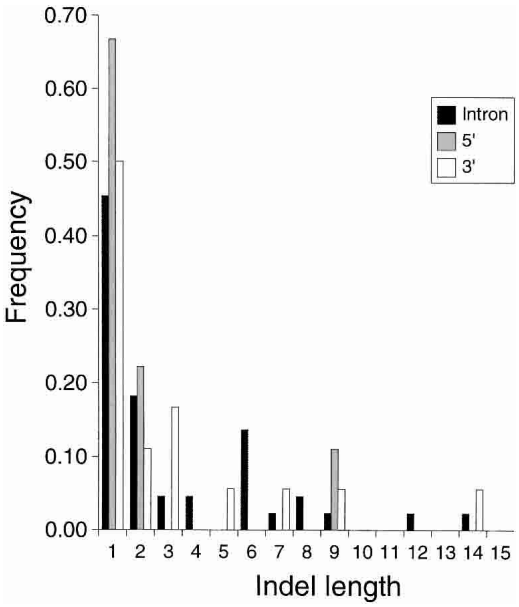

Figure 1 Frequency distribution of indel length in intronic and intergenic segments upstream (5′) and downstream (3′) from the start or stop codon, in DNA sequences of *Drosophila sechellia* and *Drosphila simulans*. Two intronic indels of length 29 and 37 have been omitted to aid clarity.

2 ln likelihood ratio $\approx \chi^2_2 = 4.5$; $P = 0.11$), although there is a suggestion that the number of indels in 5′-intergenic regions is lower in relation to the number of substitutions than in intronic DNA (Table 1; 2 ln likelihood ratio $\approx \chi^2_1 = 4.4$; $P = 0.04$, uncorrected for multiple tests).

Evolutionary Conservation in Intronic DNA Sequences of *Drosophila*

We computed estimates of the level of constraint in *D. melanogaster/simulans* intron sequences by the two lineage approach, using synonymous sites as the putatively neutral standard, as described in Methods, under the assumption that the equilibrium G + C content (f_e) is equal to the G + C content of intronic sequences in our data set (0.37). Separate estimates were made for complete intron sequences, and for intron sequences stripped of putative 5′- and 3′-splice control sequences (Table 2). Estimates of constraint in intron sequences are negative in sequences either including ($P = 0.1$) or excluding ($P = 0.007$) splice control sequences. Negative estimates of constraint imply that fourfold sites evolve more slowly than intronic sites (particularly those sites that are outside splice control regions), after differences in base composition have been accounted for. We investigated the slightly higher constraint in sequences including putative splice control sequences by calculating constraints for groups of bases at the 5′- and 3′-ends of intron sequences (Table 3). In the sequences analyzed, conservation is absolute for the 5′-splice-junction GT dinucleotide, and constraint is also strong for base pairs 3–6 at the 5′-end; all these bases are important in delimiting the 5′-end of introns (Sharp 1994). With the exception of the invariant 3′-splice-junction AT dinucleotide, constraint is nonsignificant at the 3′-end, a somewhat surprising result, considering that the consensus for nucleotides 5–16 is a run of pyrimidines showing moderate conservation across eukaryotic lineages (Sharp 1994). Constraint in intronic splice sequences, calculated using non-splice-control intronic sequences as the putatively neutral standard, gives similar results to those shown in Table 3 (i.e., 5′ bp 3–6: $C = 0.715$, SEM = 0.0484; 3′ bp 3–16: $C = 0.0517$, SEM = 0.0746).

We also calculated constraint specific to the *D. melanogaster* and *D. simulans* lineages by the three-lineage approach described in Methods, using *Drosophila yakuba* sequences as the outgroup. The results are consistent for those obtained for the two-sequence method, and suggest weak constraint in intron sequences lacking putative splice sites in both species. The level of constraint is significantly lower for bases 3–6 of the 5′-end in *melanogaster* than *simulans* ($P = 0.048$; bootstrap analysis, Table 4), lending support to the idea of a lower intensity of selection in the lineage leading to *D. melanogaster*. The constraint difference at the 3′-end (Table 4) is nonsignificant ($P > 0.8$).

Evolutionary Conservation of Intergenic DNA Sequences of *Drosophila*

We computed constraint by the two-lineage method in 5′- and 3′-intergenic DNA sequences of *D. melanogaster* and *D. simulans* using synonymous sites as the putatively neutral standard (Table 2). The estimated levels of constraint contrast with the results for intronic DNA: There is moderate to strong positive constraint in much of the 1000 bp of intergenic DNA analyzed, implying the action of purifying selection. The average levels of constraint in segments of up to 500 bp upstream and downstream of genes are 0.174 (SEM = 0.058) and 0.256 (SEM = 0.135), respectively. The corresponding values for constraint computed using intronic nucleotides excluding splice sites as the putatively neutral standard are 0.373 (SEM = 0.078) and 0.522 (SEM = 0.082). These

Table 2. Observed and Expected Numbers of Nucleotide Substitutions Along With Estimates for Constraint in Noncoding DNA Sequences of Three Categories, Calculated by the Two Lineage Approach Using Four-Fold Sites of Homologous Genes from *D. simulans* and *D. melanogaster* as a Standard

DNA category	Data set	Number of loci	Base pairs per locus	Substitutions per locus		Constraint (SEM)
				Observed (SEM)	Expected (SEM)	
Intronic	Complete	91	228	16.23 (1.69)	14.86 (1.56)	−0.094 (0.059)
Intronic	Splice sequences omitted	91	190	13.72 (1.63)	11.75 (1.45)	−0.17 (0.069)
5′ intergenic	bp 1–100	77	99	5.43 (0.36)	8.29 (0.46)	0.34 (0.053)
	bp 101–200	73	98	5.58 (0.45)	7.97 (0.46)	0.30 (0.056)
	bp 201–300	69	98	6.51 (0.53)	7.82 (0.49)	0.16 (0.079)
	bp 301–500	66	188	13.37 (1.06)	13.86 (0.86)	0.03 (0.088)
3′ intergenic	bp 1–100	42	95	5.78 (0.69)	7.29 (0.54)	0.21 (0.086)
	bp 101–200	31	88	5.35 (0.77)	6.58 (0.83)	0.17 (0.14)
	bp 201–300	21	92	5.91 (0.79)	6.65 (1.06)	0.093 (0.17)
	bp 301–500	18	150	7.10 (1.30)	11.60 (2.13)	0.36 (0.19)

Table 3. Estimates of Constraint in Runs of Nucleotides Close to the 5′- or 3′-Ends of Introns, Calculated Using the Two Lineage Method, Applied to Sequences from *D. melanogaster* and *D. simulans*

Intron end	Base pairs	Substitutions: Observed (SEM)	Substitutions: Expected (SEM)	Constraint (SEM)
5′	1–2	0 (0.0)	0.294 (0.021)	1.00 (0.00)
	3–4	0.145 (0.037)	0.277 (0.021)	0.47 (0.13)
	5–6	0.055 (0.024)	0.295 (0.022)	0.82 (0.077)
3′	1–2	0 (0.0)	0.292 (0.021)	1.00 (0.00)
	3–16	2.248 (0.202)	1.912 (0.142)	−0.178 (0.098)

findings agree with Duret and Mouchirod (2000), who found a negative correlation between gene expression level and substitution rate in 5′- and 3′-untranslated regions of mammals, indicating the operation of purifying selection.

DISCUSSION

In contrast to intron sequences in mammals and several other taxa (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002), intron sequences tend to be rather short in *Drosophila*, with a peak length of only ~60 bp (Adams et al. 2000), and several studies have revealed precursor mRNA secondary structure in intronic sequences (Stephan and Kirby 1993; Kirby et al. 1995; Leicht et al. 1995). We therefore expected that constraints would be easily observed in *Drosophila*, if introns commonly contain gene expression control sequences.

The analysis did not bear this expectation out. The results are consistent with somewhat faster evolution at most intronic sites than fourfold sites, which themselves are thought to be under weak selection. Under the assumption of a nonequilibrium model of sequence evolution, our analysis indicates that intronic sequences outside splice control sequences evolve ~17% faster than fourfold sites of adjacent genes. Similar findings have recently been reported in primates (Chen and Li 2001; Hellmann et al. 2003) and rodents (Keightley and Gaffney 2003), although a different study in primates did not reveal the pattern (Subramanian and Kumar 2003). We examined the robustness of the result by changing f_e, the equilibrium GC content. Values of f_e below 0.53 give higher rates of evolution at intronic sites than fourfold sites, and for values below 0.41 the difference in rates (as measured by constraint) is significant at $P < 0.05$. In introns, moderate to strong constraint was only detected between *melanogaster* and *simulans* at the dinucleotides at exon–intron boundaries and at 5′ nucleotides 3–6; these latter nucleotides also show a notable difference in constraint between the two lineages, possibly brought about by a difference in the long-term effectiveness of selection between the species (Aquadro et al. 1988; Akashi 1995; Moriyama and Powell 1996; Andolfatto 2001; Eyre-Walker et al. 2002). This pattern of constraint close to intron boundaries implies that mutations at these sites are slightly deleterious (Ohta 1992), and is therefore indirect evidence that the remaining sequences are genuinely evolving free from selective constraints. There is little difference between expected and observed numbers of substitutions between *melanogaster* and *simulans* (Table 4), an observation that is also consistent with models of weak selection, because selection coefficients of the order of the reciprocal of N_e are predicted to have little influence on substitution rates. The present findings are in broad agreement with McVean and Vieira (2001), who found that predicted rates of substitution were similar to observed rates in *Drosophila*. Our results concord with observations of the density of nucleotide polymorphisms in human introns as a function of distance from the 5′- or 3′-end (F.A. Kondrashov pers. comm.); there was no evidence for selection operating beyond about nucleotide 10 from the 5′- or 3′-end.

Our results contrast with recent estimates of the levels of constraint in introns and intergenic DNA in *Drosophila* (Bergman and Kreitman 2001) and *Caenorhabditis* (Shabalina and Kondrashov 1999), in which constraint was calculated from the fraction of conserved nucleotides in alignable blocks of DNA between distantly related species. Surprisingly, frequencies of conserved blocks in introns and intergenic DNA were similar to each other (of the order of 20%). However, variability in the mutation rate from region to region (Clark 2001) could give the false impression of evolutionary conservation in a segment that is evolving at the neutral rate. Furthermore, alignment of noncoding DNA is problematical with widely diverged species. Any noncoding DNA alignment that is not based on a model of indel evolution is likely to be biased (Thorne et al. 1991), and estimates of numbers of nucleotide substitutions may either be too high or too low depending on whether the alignment algorithm inserts too few or too many indels. Estimates of the proportion of conserved blocks in noncoding regions between mouse and human (Jareborg et al. 1999) are also likely to be susceptible to such biases.

The data in Table 3 indicate that the number of constrained nucleotides per intron is ~4.1. If there are 41,000 introns in the *Drosophila* genome (Adams et al. 2000), the predicted number of constrained nucleotides in introns is therefore only 0.17 Mb. The level of constraint at amino acid sites of *Drosophila* genes has been estimated to be ~84% (Eyre-Walker et al. 2002), implying that the total number of constrained amino acid sites in the *Drosophila* genome is ~16 Mb (~14,000 protein-coding genes, comprising an average of 591 codons [Adams et al. 2000], and about three-quarters of sites in coding DNA lead to an amino acid change if mutated). The number of constrained nucleotides in introns is therefore relatively small in relation to the protein-coding segment of the genome. However, the number of constrained nucleotides in intergenic DNA could potentially be of the same order as in coding DNA. For example, if we assume the average constraint values calculated relative to intronic sequences, we

Table 4. Estimates of the Level of Constraint in Introns of 38 Loci From *D. melanogaster* and *D. simulans*, Computed Using the Three-Lineage Method

Lineage	Data set	bp Per locus	Substitutions: Observed (SEM)	Substitutions: Expected (SEM)	Constraint (SEM)
melanogaster	Complete	204	7.05 (1.07)	7.17 (1.15)	0.010 (0.11)
simulans			6.49 (1.04)	7.82 (1.59)	0.15 (0.14)
melanogaster	Splice sequences omitted	161	5.87 (0.99)	5.45 (0.96)	−0.087 (0.13)
simulans			5.39 (0.95)	5.90 (1.36)	0.065 (0.15)
melanogaster	5′ bases 3–6	7.8	0.184 (0.074)	0.326 (0.045)	0.44 (0.21)
simulans			0.052 (0.036)	0.370 (0.057)	0.86 (0.10)
melanogaster	3′ bases 3–16	27.3	1.003 (0.200)	1.112 (0.196)	0.084 (0.19)
simulans			1.034 (0.200)	1.205 (0.201)	0.12 (0.21)

obtain 14,000 genes × 1000 bp × 0.44 = 6.2 Mb. This is a minimum estimate, whose value could be much larger if there are appreciable functional constraints deep in the intergenic DNA.

METHODS

Data

Homologous gene sequences (partial or complete) from *D. simulans* and *D. melanogaster*, and, where available *D. yakuba*, were compiled from GenBank. Genes were selected if they contained at least one intron, or at least 60 bp of intergenic DNA upstream or downstream for the start or stop codon. Coding sequences were aligned using CLUSTAL (Thompson et al. 1994) and corrected manually. Noncoding sequences were aligned using MCALIGN (P.D. Keightley and T. Johnson, unpubl.), a procedure that attempts to find the most probable alignment according to specific models of indel evolution (see below). The parameters of the alignment model were derived from data on relative rates of indels and nucleotide substitutions between *D. simulans* and *D. sechellia* (Table 1), and the frequency distribution of indel lengths (Fig. 1). Noncoding DNA alignments of *D. simulans* and *D. sechellia* are virtually unambiguous, by alignment with any standard heuristic alignment method. The most probable alignment of the *D. simulans/melanogaster/yakuba* sequences were used in subsequent analysis. Intergenic DNA was categorized either as 5′ or 3′. In cases in which genes are so close together in the genome that this categorization was ambiguous, stretches of DNA were arbitrarily assigned to the 5′ category, although they could have been considered to belong to the 3′-segment of an adjacent gene. Intergenic DNA includes any DNA that is 5′ or 3′ from the start or stop codon, and therefore contains elements of transcribed untranslated DNA.

Introns were either analyzed as complete sequences or partial sequences after removal of putative splice control sequences. The base pairs removed were 1–6 at the 5′-end and 1–16 at the 3′-end. The exact limits of the control sequences are somewhat arbitrary (Sharp 1994).

Lists of loci are shown in Supplemental Tables 1 and 2, and aligned sequences are available from PDK's Web site.

Sequencing of Additional *Drosophila simulans* Intergenic Sequences

We obtained additional intergenic DNA sequences from *Drosophila simulans* by sequencing the 5′-flanking regions of genes for which the orthologous coding sequences were available for both *simulans* and *melanogaster* on GenBank. Genes for which there was only a short length of available coding sequence in *simulans* were excluded (we used an arbitrary cutoff of 200 bp), and we did not sequence upstream DNA from previously sequenced *simulans* loci. Primers for sequencing were designed (using Primer Premier 5.00; Premier Biosoft International) to be ~650 to 700 bp apart, based on the *melanogaster* sequence. Upstream primers were usually designed from the noncoding *melanogaster* sequence (where possible an upstream coding sequence was used), and downstream primers were designed using the *simulans* coding sequence.

Genomic DNA for PCR reactions was prepared (Gentra Systems, Research Triangle Park) from a single partially inbred male *Drosophila simulans* fly collected in Aswan, Egypt in 2001. A single male fly was used as a source of DNA in all cases to reduce sequencing problems associated with heterogeneity in template DNA. A combination of standard PCR and asymmetric PCR (Miller et al. 2003) was used to amplify the appropriate section of DNA. If the primers failed to amplify the appropriate section of DNA, the primers were redesigned. If the appropriate DNA segment still could not be amplified after the primers had been redesigned three times, investigation of the gene was terminated. In 18 cases out of 63, we could not obtain sufficient amplification of the appropriate section of DNA.

Purified PCR products were sequenced on both strands using an ABI prism BigDye terminator cycle sequencing kit (Applied Biosystems) and run on an Applied Biosystems 3730 DNA Analyzer (Applied Biosystems). Sequences from each strand for each gene were then assembled using Sequencher 3.0 software (Gene Codes), and alignments were checked manually. The GenBank accession numbers are AY459538–AY459582.

Likelihood Ratio Test for Variation in Rates of Indels Relative to Nucleotide Substitutions

The test was constructed under the assumption that sequences are sufficiently closely related such that multiple hits can be ignored, and that the number of indels is linearly related to the number of nucleotide substitutions. Assume that there are n categories of noncoding DNA (say, $n = 3$ with 1 = intronic, 2 = 5′-intergenic, and 3 = 3′-intergenic). Let k_i be a parameter for the fraction of nucleotide differences between sequences of category i, and $\theta_i k_i$ be a compound parameter for the fraction of indels differentiating sequences in category i. Under the assumption of independent binomially distributed nucleotide substitution and indel numbers, the likelihood of observing n_i substitutions and g_i indels is

$$L_i = k_i^{n_i} (1 - k_i)^{l_i - n_i} (\theta_i k_i)^{g_i} (1 - \theta_i k_i)^{m_i}, \quad (1)$$

where l_i is the number of base pairs in the sequence (excluding indels) and m_i is the number of sites not occupied by an indel. The likelihood of the observations of three categories of DNA is $L = L_1 \times L_2 \times L_3$, and the two models are compared according to $\theta_1 \neq \theta_2 \neq \theta_3$ (full model), and $\theta_1 = \theta_2 = \theta_3$ (reduced model). The likelihood with respect to k and θ was maximized numerically.

Alignment of Noncoding DNA Sequences According to Models of Indel Evolution

Alignment was carried out by a procedure MCALIGN, available at PDK's Web pages. The procedure uses a Monte Carlo algorithm to search for the most probable alignment of a pair of sequences or of three sequences that includes an outgroup, based on a model of indel evolution. The parameters of the model are θ, the rate of indels relative to nucleotide substitutions, and a vector parameter w specifying the frequency distribution of indel lengths. Because θ is a parameter of the model, estimated alignments containing large (small) numbers of nucleotide differences tend to have large (small) numbers of indels, a pattern supported by mouse–human sequence alignments (Hardison et al. 2003). In aligning pairs of sequences, the Jukes-Cantor method is used to correct for multiple nucleotide substitutions. For three sequences, parsimony is used to assign substitutions and indel events to the ingroup or the outgroup, and the probability of the alignment is the product of probabilities for the ingroup and outgroup.

The model parameters θ and w come from external data, in the present case from alignments of noncoding DNA of *D. simulans* and *D. sechellia*. Values of θ from Table 1 were used to parameterize three alternative models, for aligning intronic, 3′-intergenic, or 5′-intergenic DNA. The vector parameter w was assumed to be the same for each model, and was derived from the frequency distribution of indels in introns, after some smoothing of the distribution had been applied.

Two-Lineage Approach to Compute Constraint in Noncoding DNA

Following distance-based methods for calculating constraint in coding DNA (Eyre-Walker and Keightley 1999), the present method uses rates of substitution at fourfold sites or other putative neutral sites of a gene to predict expected numbers of substitutions in an adjacent noncoding DNA segment, such as an intron or flanking sequence, assuming equal rates of mutation in the sequences. The method takes into account differences in base composition. The expected numbers of substitutions (E) are compared with the observed numbers (O) to calculate constraint (C). For example, if $E = O$, the constraint in the noncoding segment is

zero; if $O = 0$, constraint takes the value of 1. The method is only applicable to closely related species for which multiple hits can be safely ignored.

In a pairwise comparison, it is not possible to determine the direction of a particular substitution (i.e., whether a C↔T difference is caused by a C→T or a T→C substitution). However, it is possible to infer the proportion of changes that are in a particular direction if we assume or know the equilibrium base composition. Let us group Gs and Cs together, and As and Ts together. Let f_e be the equilibrium G+C content of the sequence; this is the G+C content that the sequence will eventually reach, and let z be a mutation rate parameter such that the rate at which A or T sites change to G or C is zf_e, and the rate at which G or C sites change to A or T is $z(1 - f_e)$. We can then use the present and equilibrium G+C content to infer the proportion of observed AT↔GC differences that go in a particular direction (this category of differences involves the following pairwise differences: A↔G, A↔C, T↔G, and T↔C). However, with only two species, we cannot infer whether an observed G↔C difference is caused by a G-to-C mutation or C-to-G mutation (this would require a parsimony approach). Similarly, we cannot assign polarity to any observed A↔T differences. We can therefore only calculate four different rates ($i = 1 \ldots 4$), two pairwise rates (A↔T and G↔C) and two directional rates (AT→GC and GC→AT). If we consider evolution over a fairly short period of time so that the G+C content does not change dramatically (or not at all if the sequence is at equilibrium), then the numbers (X) of AT→GC mutations and GC→AT mutations are

$$X_{(AT\rightarrow GC)} = f_e z N(1 - f_a) \quad (2)$$

$$X_{(GC\rightarrow AT)} = (1 - f_e) z N f_a \quad (3)$$

where f_a is the G+C content of the sequence being considered at the separation of the two species being considered (in practical terms we can assume that this is equal to the present G+C content if the time of divergence is small). Under these assumptions, an equilibrium is reached when the number of mutations in one direction equals the number of mutations in the reverse direction, that is, when $f_e z N(1 - f_a) = (1 - f_e) z N f_a$, or when $f_a = f_e$.

The total number of AT↔GC differences that have occurred ($X_{AT\leftrightarrow GC}$) can be written as the sum of equations 2 and 3, and rearranging this for z:

$$z = \frac{X_{(AT\leftrightarrow GC)}}{((1 - f_e)f_a + f_e(1 - f_a))N} \quad (4)$$

By substituting equation 4 into equations 2 and 3, we can remove the unknown parameter z and express the two estimates of the number of directional mutations in terms of f_e, f_a, and the number of pairwise AT↔GC differences. If we then divide by the number of sites at which each of these types of mutation could occur [$N(1 - f_a)$ and Nf_a, respectively], we can obtain an estimate of the per site rate of AT→GC and GC→AT mutations.

$$K_{(AT\rightarrow GC)} = \frac{X_{(AT\leftrightarrow GC)} f_e}{((1 - f_e)f_a + f_e(1 - f_a))N} \quad (5)$$

$$K_{(GC\rightarrow AT)} = \frac{X_{(AT\leftrightarrow GC)}(1 - f_e)}{((1 - f_e)f_a + f_e(1 - f_a))N} \quad (6)$$

The predicted (expected) number of substitutions in the noncoding DNA segment is,

$$E = \sum_{i=1}^{4} K_i M_i, \quad (7)$$

where M_i is the number of sites in the noncoding segment corresponding to rates of type i. The observed number of differences in the segment, O, is the number of nucleotide differences in the noncoding segment. Constraint for a segment is given by $C = 1 - O/E$, or, for several segments it is

$$C = 1 - \sum O_j / E_j, \quad (8)$$

where the summation is carried out over segments. Standard errors of O, E, and C are calculated by bootstrapping the data, by gene (Eyre-Walker and Keightley 1999).

Three-Lineage Approach to Compute Constraint in Noncoding DNA

The basis of the approach is to calculate rates for all possible kinds of nucleotide substitutions at fourfold and twofold degenerate synonymous sites of a gene, and to use these rates to calculate expected numbers of substitutions in an adjacent noncoding DNA segment. The numbers of substitutions at synonymous and noncoding sites are estimated using parsimony. The rates for the 12 possible kinds of synonymous substitution in one of the branches, $K_{A\rightarrow T}$, $K_{A\rightarrow C}$, $K_{A\rightarrow G}$, $K_{T\rightarrow A}$, and so on, are computed by taking weighted averages of the fraction of differences at fourfold, and, where applicable, twofold degenerate sites. Under the assumption of neutral evolution at synonymous sites and equal mutation rates in the coding and noncoding DNA, the expected number of substitutions in the noncoding segment associated with gene i is, therefore,

$$E_i = N_A(K_{A\rightarrow T} + K_{A\rightarrow C} + K_{A\rightarrow G}) + N_T(K_{T\rightarrow A} + \ldots) + N_C(K_{C\rightarrow A} + \ldots) + N_G(K_{G\rightarrow A} + \ldots) \quad , \quad (9)$$

where N_A, N_T, N_C, and N_G, are the numbers of A, T, C, and G nucleotides, respectively, in the noncoding sequence. In cases in which all three base pairs differ, averages of $K_{A\rightarrow T} + K_{A\rightarrow C}$, and so on, are calculated, weighted by the probabilities of alternative ancestral states, on the assumption that the relative lengths of the branch from the *melanogaster*/*simulans* common ancestor (a) to *simulans* and *melanogaster* are 0.2, and the relative length of the branch from a to *yakuba* is 0.6. This model gives probabilities for the ancestral state being the *melanogaster* or *simulans* base of 0.462, and for the *yakuba* base of 0.0769. O_i is the observed number of substitutions in the aligned noncoding sequence associated with gene i. The average constraint is calculated by equation 8.

ACKNOWLEDGMENTS

We thank Brian Charlesworth, Toby Johnson, Penny Haddrill, and two reviewers for helpful comments and suggestions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. *Genetics* **139:** 1067–1076.

———. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144:** 1297–1307.

———. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151:** 221–238.

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* **18:** 279–290.

Aquadro, C.F., Lado, K.M., and Noon, W.A. 1988. The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans* .1. contrasting levels of naturally-occurring DNA restriction map variation and divergence. *Genetics* **119:** 875–888.

Begun, D.J. and Whitley, P. 2002. Molecular population genetics of *Xdh* and the evolution of base composition in *Drosophila*. *Genetics* **162:** 1725–1735.

Bergman, C.M. and Kreitman, M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11:** 1335–1345.

Charlesworth, B. and Charlesworth, D. 1998. Some evolutionary consequences of deleterious mutations. *Genetica* **103:** 3–19.
Chen, F.-C. and Li, W.-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.*Am. J. Hum. Genet.* **68:** 444–456.
Clark, A.G. 2001. The search for meaning in noncoding DNA. *Genome Res.* **11:** 1319–1320.
Duret, L. and Mouchiroud, D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17:** 68–74.
Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.
Eyre-Walker, A. and Bulmer, M. 1995 . Synonymous substitution rates in enterobacteria. *Genetics* **140:** 1407–1412.
Eyre-Walker, A. and Keightley, P.D. 1999. High genomic deleterious mutation rates in hominids. *Nature* **397:** 344–347.
Eyre-Walker, A., Keightley, P.D., Smith, N.G.C., and Gaffney, D. 2002. Quantifying the slightly deleterious model of molecular evolution. *Mol. Biol. Evol.* **19:** 2142–2149.
Gilbert, W. 1978. Why genes in pieces. *Nature* **271:** 501.
Glazko, G.V., Koonin, E.V., Rogozin, I.B., and Shabalina, S.A. 2003. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* **19:** 119–124.
Gu, X. and Li, W-H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40:** 464–473.
Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.
Hellmann, I., Zollner, S., Enard W., Ebersberger, I., Nickel, B., and Pääbo, S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13:** 831–837.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.
Keightley, P.D. and Gaffney, D.J. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci.* **100:** 13402–13406.
Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK.
Kirby, D.A., Muse, S.V., and Stephan, W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc. Natl. Acad. Sci.* **92:** 9047–9051.
Kondrashov, A.S. 1995. Contamination of the genome by very slightly deleterious mutations. Why have we not died 100 times over? *J. Theor. Biol.* **175:** 583–594.
Kondrashov, A.S. and Crow, J.F. 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2:** 229–234.
Leicht, B.G., Muse, S.V., Hanczyc, M., and Clark, A.G. 1995. Constraints on intron evolution in the gene encoding the myosin alkali light-chain in *Drosophila*. *Genetics* **139:** 299–308.
Li, W.-H. 1997. *Molecular evolution.* Sinauer, Sunderland, MA.
Li, W.-H. and Graur, D. 1991. *Fundamentals of molecular evolution.* Sinauer, Sunderland, MA.
Ludwig, M.Z. and Kreitman, M. 1995. Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Mol. Biol. Evol.* **12:** 1002–1011.
McVean, G.A.T. and Vieira, J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157:** 245–257.
Miller, M.D., Duan, S., Lovins, E.G., Kloss, E.F., and Kwok, P. 2003. Efficient high-throughput resequencing of genomic DNA. *Genome Res.* **13:** 717–720.
Moriyama, E.N. and Powell, J.R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13:** 261–277.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Ann. Rev. Ecol. Syst.* **23:** 263–286.
Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.
Shabalina, S.A. and Kondrashov, A.S. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74:** 23–30.
Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A., and Kondrashov, A.S. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17:** 373–376.
Sharp, P.A. 1994. Split genes and RNA splicing. *Cell* **77:** 805–815.
Shields, D.C., Sharp, P.M., Higgins, D.G., and Wright, F. 1988. Silent sites in *Drosophila* genes are not neutral—Evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5:** 704–716.
Stephan, W. and Kirby, D.A. 1993. RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. *Genetics* **135:** 97–103.
Subramanian, S. and Kumar, S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13:** 838–844.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thorne, J.L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114–124.
———. 1992. Inching toward reality—An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:** 3–16.

WEB SITE REFERENCES

http://homepages.ed.ac.uk/eang33/; executables, source code, and user instructions for MCALIGN.
http://homepages.ed.ac.uk/eang33/; aligned sequence data files.

Received March 11, 2003; accepted in revised form December 1, 2003.